同济大学

毕业设计(论文)任务书

（适用于工科类、理科类专业）

课题名称	整合核小体定位信息预测非典型DNA位点
副标题	
学院（系）	生命科学与技术学院
专　业	生物信息
学生姓名	张佳胜　　学　号　　1753096

毕业设计(论文)起讫时间：

自 2020 年 9 月 14 日至 2021 年 6 月 20 日共 32 周

指导教师签名　　年　月　日

教学院长(系主任)签名　　年　月　日

一、 毕业设计（论文）的课题背景

课题来源：自拟

主要背景介绍：

四链 DNA(four-stranded DNA)结构，被称为 G-quadruplexes(G4s)，又叫鸟嘌呤(guanine)四链体，一般产生于 G 四平方面的自我堆叠[1]。过往的研究过程中，G4s 通常发现于体外(vitro)，而近来的研究以及相关证据告诉我们，这种结构也存在于生物体内(vivo)[2]，通过实验发现，G4 的位点不随机，具有特定的序列模式，与部分关键性基因区域（如启动子，端粒以及基因区域的第一个内含子等）相关[3]。一方面， G4s 在体内承担着某些潜在的重要功能；另一方面，多项研究结果表明，G4s 与 DNA 的复制，转录、基因表达、端粒的稳定性[2-3]以及表观遗传记忆[4]等都具有一定的相关性。同时，更进一步的病理学研究发现，探究 G4s 的作用以及特异性可以帮助进行肿瘤学的研究[1]，具有极大的医疗潜力。综上可见，G4s 结构的相关研究得到了生物研究领域的广泛关注，具有一定的研究价值。

为了研究 G4s，研究者们需要定位其在基因组中的位置。而现在可以通过染色体免疫共沉淀然后测序的技术(G4 ChIP-seq)[1]等方式进行 G4s 位点的探测，但它们仍然存在一定的缺陷，这导致了部分的 G4s 结构位点尚不清楚，相应的实验数据有限。研究者们根据之前发现的 G4s 的特定序列模式，尝试在计算机层面针对全基因组的 G4s 位点进行大通量的预测[7]，开发出了一些算法和工具。现有的主要算法和工具可以分为三大类：通过 G4 序列模式进行正则匹配、滑动窗口结合打分矩阵[8-10]以及机器学习[13]的方式。前两者基于实验经验（包括核小体的特定序列模式，特殊结构位点等）进行预测，后者则完全由数据驱动。但是，根据匹配模式的宽泛程度以及打分的形式，不同的算法都具备各自的优点和缺点，现有的算法也很难对不同细胞中的 G4s 结构位点进行准确预测。故而，我们需要引入一种新的算法模型，使得 G4s 的预测准确性得到提升，并且能够解析不同组织细胞中的 G4s 结构异质性。

核小体是由 DNA 和组蛋白形成的染色质基本结构单位，它将链状 DNA 通过组蛋白结合缠连的方式组合成一个结构单元。研究者发现，核小体定位与 G4s 位点存在一定的相关性，这启发我们借助核小体定位信息来进行模型的构建。现有的核小体定位具备一套比较完备的流程以及数据质量检测的方法，我们能够拿到大量不同细胞类型中比较完备的全基因组核小体定位数据，这能够帮助我们解析不同细胞类型中的 G4s 结构位点图谱，预测其位点。

基于上述背景，本次课题将从以下几个方面展开：首先，我们期望能够通过我们的整合核小体数据辅助序列模式匹配来构建一个预测模型，实现预测功能性以及准确性的提升；其次，我们会整理现有的不同算法，比较它们的预测效果，并提出我们的模型优势；最后，我们可能还会对模型进行验证，评估模型的可靠性。

二、毕业设计（论文）的技术参数（研究内容）

目标与需要达到的指标：

初步的核小体定位数据以及 G4 ChIP-seq 数据的处理以及分析

证明核小体定位与 G4 位点的相关性

预测模型构建以及算法优化

整理现有方法同时针对这些模型进行比较评估

验证模型的准确性与可靠性

主要内容：

项目初期，我们搜集并预处理核小体定位数据以及 G4 ChIP-seq 的 G4s 位点数据，分析其相关性；然后，我们根据数据之间的关联设计模型和算法，并整理现有的一些 G4s 预测方法，进行预测模型的比较和评估，指出我们模型的特色与优势；最后，我们可能会利用一部分 G4s 位点数据来验证模型的准确性和可靠性。我们的期望是，我们的模型能够针对不同细胞的 G4s 结构位点进行预测，并且相比较于其他方法具备较好的预测效果或者更多优势。

三、毕业设计（论文）应完成的具体工作

1. 项目文献收集、核小体定位数据以及 G4 ChIP-seq 数据搜集（2020.10.15-2020.10.30）
2. 针对上述的原始数据进行整理以及预处理（2020.10.15-2020.10.30）
3. 核小体定位与 G4 位点的相关性探究（2020.10.31-2020.11.31）
4. 开题报告与综述（2020.10.31-2020.12.30）
5. 算法调研、比较以及选取（2020.11.31-2020.12.15）
6. 利用核小体定位结合已有的 motif 算法构建预测模型（2020.12.15-2020.01.30）
7. 比较模型的预测效果以及指出算法的优势（2020.03.01-2021.03.30）
8. 预测组织特异性的 G4 结构位点图谱（2021.03.31-2021.04.20）
9. 验证模型的准确性和可靠性(2021.04.20-2021.05.09)
10. 论文撰写与毕业答辩（2021.05.9-2021.06.20）

四、毕业设计（论文）进度安排

序 号	设计（论文）各阶段名称	时间安排（教学周）
	毕设动员，题目征集双选。学生进实验室与指导老师确定研究任务，完成任务书撰写。	2020.9.14-2020.10.15（第 1-4 周）
	开展研究工作，进行文献调研以及数据的收集，数据预处理	2020.10.15-2020.10.30（第 5-8 周）
	撰写提交开题报告，情况表 核小体定位与 G4 位点的相关性探究	2020.10.31-2020.11.30（第 9-12 周）
	撰写综述及文献翻译 算法调研、比较以及选取	2020.12.1-2020.12.30（第 13-16 周）
	实验，进度自查，中期检查	2020.12.31-2021.5.8

		（第 17-28 周）
	核小体定位结合已有的 motif 算法构建预测模型	2020.12.15 - 2020.1.30 （第 15-19 周）
	比较模型的预测效果 指出算法的优势	2020.3.1-2021.3.30 （第 20-23 周）
	预测组织特异性的 G4 结构位点图谱	2021.3.31-2021.4.20 （第 24-26 周）
	验证模型的准确性和可靠性	2021.4.20-2021.5.9 （第 27-28 周）
	论文撰写，毕业答辩	2021.5.9-2021.6.20 （第 29-34 周）

同组学生姓名：无

五、应收集的资料及主要参考文献

[1] Hänsel-Hertsch R, Di Antonio M, Balasubramanian S. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential [J]. Nat Rev Mol Cell Biol. 2017 May;18(5):279-284.

[2] Bochman ML, Paeschke K, Zakian VA. DNA secondary structures: stability and function of G-quadruplex structures [J]. Nat Rev Genet. 2012 Nov;13(11):770-80.

[3] Rhodes D, Lipps HJ. G-quadruplexes and their regulatory roles in biology [J]. Nucleic Acids Res. 2015 Oct 15;43(18):8627-37.

[4] Mukherjee AK, Sharma S, Chowdhury S. Non-duplex G-Quadruplex Structures Emerge as Mediators of Epigenetic Modifications [J]. Trends Genet. 2019 Feb;35(2):129-144.

[5] Chambers VS, Marsico G, Boutell JM, et al. High-throughput sequencing of DNA G-quadruplex structures in the human genome [J]. Nat Biotechnol. 2015 Aug;33(8):877-81.

[6] Hänsel-Hertsch R, Beraldi D, Lensing SV, et al. G-quadruplex structures mark human regulatory chromatin [J]. Nat Genet. 2016 Oct;48(10):1267-72.

[7] Puig Lombardi E, Londoño-Vallejo A. A guide to computational methods for G-quadruplex prediction [J]. Nucleic Acids Res. 2020 Jan 10;48(1):1-15.

[8] Brázda V, Kolomazník J, Lýsek J, et al. G4Hunter web application: a web server for G-quadruplex prediction [J]. Bioinformatics. 2019 Sep 15;35(18):3493-3495.

[9] Brazda V, Kolomaznik J, Mergny JL, et al. G4Killer web application: a tool to design G-quadruplex mutations [J]. Bioinformatics. 2020 May 1;36(10):3246-3247.

[10] Hon J, Martínek T, Zendulka J, et al. pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R [J]. Bioinformatics. 2017 Nov 1;33(21):3373-3379.
[11] Yang D. G-Quadruplex DNA and RNA [J]. Methods Mol Biol. 2019;2035:1-24.
[12] Kwok CK, Merrick CJ. G-Quadruplexes: Prediction, Characterization, and Biological Application [J]. Trends Biotechnol. 2017 Oct;35(10):997-1013.
[13] Garant JM, Perreault JP, Scott MS. G4RNA screener web server: User focused interface for RNA G-quadruplex prediction [J]. Biochimie. 2018 Aug;151:115-118.